Workloads Table for the paper: A Strawman Model for Architecting DRAM Caches

I. Experimental Methodology

A. Simulation Setup

We evaluate our proposed design enhancements using the GEM5 [11] simulator integrated with detailed DRAM-Cache and PCM simulators. The DRAM-Cache simulator faithfully models both the memory controller as well as the DRAM as a cache with accurate timing. It takes into account the cost of tag look up at the LLSC, as well as detailed DRAM organizational, and operational aspects such as DIMMs, ranks, banks, command scheduling, refreshes, activations, precharges, and column accesses as well as detailed timing parameters such as t_{FAW}. The PCM simulator models similar PCM attributes including asymetric read versus write timings, as well as differences in timing associated with SET and RESET operations during writes. The simulator implements PRIORITY-FR-FCFS scheduling with *critical sub-block first* scheme. The simulator also implements tracking of dirty bits on a per CPU cache block granularity to minimize write-back traffic.

Each program in the workload is executed in fast-forward mode for 9 billion instructions, and then in detailed cycle-accurate mode for 250 million instructions. Multi-core simulations are run until all the programs complete 250 million instructions.[1] As is the standard practice, programs that finish early continue to execute but the performance of only the first 250 million instructions is considered for each core.

The baseline machine configuration used in our studies is shown in Table I.

Processor	3.2 GHz OOO Alpha ISA
L1I Cache	32kB private, 64 byte blocks Direct-mapped, 3 cycle hit latency
L1D Cache	32kB private, 64 byte blocks, 2-way set-associative, 3 cycle hit latency
L2 Cache	For 1/4/8/16 cores: 1MB/4MB/8MB/16MB 4-way/8-way/16-way/32-way 32/128/256/512 MSHRs 64-byte blocks, 15 cycle hit latency Additional 15 cycles for tag lookup
DRAM-Cache	Stacked on-chip; 64-bit data path 256M/512M/1GB for 4/8/16 cores 8-Way Set Associative Cache Default Block Size of 64 bytes
DRAM	BL=8, IO: 1GHz, CL-nRCD-nRP=9-9-9 a rank comprises 4 x16 devices, each device has 8 banks, each bank has 4096 rows, 1024 columns
PCM	4GB/8GB/16GB for 4/8/16 cores BL=8, IO: 1GHz, CL=9, Read=50ns Write(SET)=400ns, Write(RESET)=100ns a rank comprises 4 x16 devices, each device has 8 banks, each bank has 65536 rows, 1024 columns
PCM and DRAM controllers	For 1/4/8/16 cores: 1/1/2/4 256 entry command queue PRIORITY_FR_FCFS scheduling Critical Sub-Block First Address-interleaving: rank-bank-row-column

TABLE I
CMP configuration

B. Workload and Metrics

We use multi-programmed workloads comprising programs from SPEC [12] 2000 and SPEC 2006 suites to evaluate our proposal. The workloads are typically a mix of programs with varying levels of memory intensity (based on their L2 MPKI). The workload mix used in our studies is presented in Table II.

We use the Average Normalized Turnaround Time (*ANTT* [10]) and Weighted Speedup (*WS* [13]) metrics to summarize performance. *ANTT* is defined:

$$ANTT = \frac{1}{n}\sum_{i}\frac{C_i^{shared}}{C_i^{alone}} \quad (1)$$

C_i^{alone} and C_i^{shared} refer to the CPU cycles consumed by the i^{th} program when running alone and as part of a n core workload, respectively.

WS is defined:

$$WS = \sum_{i}\frac{IPC_i^{shared}}{IPC_i^{alone}} \quad (2)$$

[1]Although we run only 250M Instructions per core in cycle-accurate mode, our 4, 8 and 16-core simulations each runs for a total of 1 Billion - 4 Billion instructions in cycle-accurate mode.

Quad-Core Workloads
Q1:(462,459,470,433),Q2:(429,183,462,459),Q3:(429,462,471,464),
Q4:(470,437,187,300),Q5:(459,464,183,433),Q6:(471,473,171,175),
Q7:(173,178,177,172),Q8:(179,188,164,191),Q9:(434,435,437,171),
Q10:(444,445,459,462),Q11:(401,410,178,177),Q12:(171,181,464,465)
Q13:(464,450,465,471),Q14:(453,433,459,410),Q15:(462,471,254,186)
Q16:(462,191,433,437),Q17:(401,473,435,177),Q18:(416,429,454,175)
Q19:(254,172,178,188)
Eight Core Workloads
E1:(462,459,433,456,464,473,450,445),
E2:(300,456,470,445,179,464,473,450),
E3:(168,183,437,401,450,435,445,458),
E4:(187,172,173,410,470,433,444,177),
E5:(434,435,450,453,462,471,164,186),
E6:(181,473,401,172,177,178,179,435),
E7:(437,459,445,454,456,465,171,197),
E8:(429,416,433,454,464,435,444,458)
E9:(183,462,453,471,473,433,254,168),
E10:(300,173,178,187,188,191,410,171),
E11:(470,177,168,434,410,172,464,171),
E12:(459,473,444,453,450,197,175,164),
E13:(471,462,186,254,462,444,410,179),
E14:(187,470,401,416,433,437,456,454),
E15:(300,459,462,470,433,172,191,471),
E16:(183,473,401,435,188,434,164,437)
Sixteen Core Workloads
S1:(462,459,433,179,183,473,450,445,444,470,429,171,168,172,435,458)
S2:(401,433,434,435,444,445,450,300,459,470,471,473,171,181,179,183)
S3:(178,177,168,172,173,187,191,410,429,434,462,473,465,458,464,445)
S4:(186,454,458,482,181,429,255,254,178,197,179,187,173,401,410,437)

TABLE II
WORKLOADS